Applied Energy 232 (2018) 197–210

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Random forest regression for online capacity estimation of lithium-ion batteries

Yi Li[a,*], Changfu Zou[b,*], Maitane Berecibar[a], Elise Nanini-Maury[c], Jonathan C.-W. Chan[d], Peter van den Bossche[a], Joeri Van Mierlo[a], Noshin Omar[a]

[a] Department of Mobility, Logistics and Automotive Technology Research Centre, Vrije Universiteit Brussel, Pleinlaan 2, Brussels 1050, Belgium
[b] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg 41296, Sweden
[c] ENGIE LAB Laborelec, Rodestraat 125, B-1630 Linkebeek, Belgium
[d] Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

HIGHLIGHTS

- Random forest regression is proposed for on-line battery capacity estimation.
- The estimation is developed from partial charging voltage-capacity data.
- Two features indicative of battery capacity fade are extracted from charging curves.
- An incremental capacity analysis is used for assisting battery feature selection.

ARTICLE INFO

Keywords:
Lithium-ion battery
On-line capacity estimation
State of health
Random forest regression
Incremental capacity analysis

ABSTRACT

Machine-learning based methods have been widely used for battery health state monitoring. However, the existing studies require sophisticated data processing for feature extraction, thereby complicating the implementation in battery management systems. This paper proposes a machine-learning technique, random forest regression, for battery capacity estimation. The proposed technique is able to learn the dependency of the battery capacity on the features that are extracted from the charging voltage and capacity measurements. The random forest regression is solely based on signals, such as the measured current, voltage and time, that are available onboard during typical battery operation. The collected raw data can be directly fed into the trained model without any pre-processing, leading to a low computational cost. The incremental capacity analysis is employed for the feature selection. The developed method is applied and validated on lithium nickel manganese cobalt oxide batteries with different ageing patterns. Experimental results show that the proposed technique is able to evaluate the health states of different batteries under varied cycling conditions with a root-mean-square error of less than 1.3% and a low computational requirement. Therefore, the proposed method is promising for online battery capacity estimation.

1. Introduction

Lithium-ion batteries (LIBs) have been widely applied as energy storage systems, such as the fields of electrified vehicles and power grids. The biggest concern about these batteries is their limited lifetime, as their performance deteriorates with usage. To prolong a battery's longevity while ensuring reliability over the entire service life, accurate diagnosis of the state of health (SOH) in real-time is essential. The SOH reflects the current capability of a battery to store and supply energy relative to that at the beginning of its life and is an indicator to evaluate the degradation level of batteries. Quantitatively, it can be calculated by the ratio of the actual cell capacity to its initial value.

Extensive research efforts have been dedicated to SOH monitoring since the last decades, resulting in different online estimation methods. These SOH monitoring techniques can be categorized into two types, namely electrical model-based and data-driven approaches. Electrical models use passive electrical components, such as resistors and capacitors, to simulate the behavior of a battery. Enabled by these models, recursive Bayesian state estimation algorithms (such as the extended Kalman filter, EKF) [1,2] and particle filter (PF) [3,4] have been

* Corresponding authors.
E-mail addresses: li.yi@vub.be (Y. Li), changfu.zou@chalmers.se (C. Zou).

https://doi.org/10.1016/j.apenergy.2018.09.182
Received 27 February 2018; Received in revised form 18 September 2018; Accepted 23 September 2018

adopted to identify and update SOH related model parameters, e.g., capacity and internal resistance, according to data acquired during operation. Most of these filters were implemented in a joint/dual configuration for the co-estimation problem of the state of charge (SOC) and capacity. For example, Plett [1] pioneered the concurrent use of two EKFs for SOC and SOH estimation. At each time step, the results of each EKF were calculated separately and then fed into its counterpart. Zou et al. [2] proposed a general multi-timescale estimation algorithm with rigorous stability analysis and then applied it to estimate battery SOC and SOH. To further improve the estimation accuracy, Schwunk et al. [3] introduced PF for battery state estimation. Although the class of electrical model-based methods can effectively estimate the capacity under certain conditions, these techniques suffer from intensive computation required by a large number of matrix operations, thereby hindering their real-time application in battery management systems (BMSs). Data-driven methods rely on a significant amount of statistical data to predict battery ageing behavior. Because physical insights and mathematical models with a set of parameters are not needed, these methods have potential to significantly reduce the computational time in comparison with electrical model-based approaches.

One of the most widely used data-driven techniques is incremental capacity/differential voltage (IC/DV) analysis. The IC/DV analysis has proven to be a powerful tool for battery capacity estimation [5]. IC is calculated by differentiating the change in battery capacity to the change in terminal voltage during either charging or discharging, while DV is defined as the inverse of IC. With this method, the voltage plateaus on charging/discharging curves can be transformed into clearly identifiable peaks on IC/DV curves. Each peak of the curve represents a specific electrochemical process taking place in the cell and can be characterized by features such as the intensity and position [6]. These peak features are closely related to battery capacity fade and can therefore be used as indicators for the SOH estimation. Weng et al. [7] estimated the battery SOH by relating it to the peak intensity of IC curves. Li et al. [8] established a linear regression relationship between battery capacity and the peak position on IC curves. However, IC/DV curves are sensitive to measurement noise inherent in battery systems [8,9]. Accordingly, proper smoothing methods have to be proposed for obtaining smooth curves that facilitate identification and evaluation of IC/DV curve features.

In addition to the IC/DV analysis, a wealth of machine learning techniques have been devised for battery SOH estimation, such as artificial neural network (ANN) [10,11], support vector machine (SVM) [12,13], regressive vector machine (RVM) [14] and Gaussian process regression (GPR) [15,16]. These SOH estimators are trained until they learn the complex mapping from the feature space to the capacity measurement space. To estimate battery SOH accurately, a critical step in machine learning algorithms is to process the data, such as measured current, voltage, and temperature, to effectively extract representative and necessary features of the battery ageing process. In general, these features can be categorized into: internal features, processed external features, and direct external features.

In details, the internal features, like battery internal resistance, capacitance and battery SOC, cannot be measured directly from BMS sensors and must resort to parameter/state estimation algorithms. Pan et al. [17] developed an extreme learning-machine-based method for battery capacity estimation, in which parameters of an electrical model, i.e., internal ohmic resistance and polarization resistance, were considered as the input data. Then, they identified these model parameters in real-time using a recursive least square algorithm. In comparison, the processed external features, e.g., peak position and intensity, are extracted from differential charging curves, like IC/DV curves [18,19] and voltage gradient curves [10,16]. Berecibar et al. [19] estimated cell capacity with a selection of features from IC/DV curves by using three different regression methods, namely linear regression, ANN and SVM. Similar work has been conducted by Wang et al. [18] with the aid of GPR. Wu et al. [10] trained a polynomial neural network based on the arc length and curvature from voltage velocity curves, and established the relation between the geometric properties of charging curves and the battery capacity. Different to the above two types of features, the direct external features are directly recorded in BMSs during battery operation, typically including the measured terminal voltage, current, and surface temperature. Hu et al. [14] applied an RVM algorithm to learn the complex dependency of the battery capacity on characteristic features extracted from voltage and current measurements during charging operation. Recently, Richardson et al. [15] proposed a capacity estimation algorithm by using GPR based on a small portion of charging voltage–time data under a constant current. They selected a subset of the smoothed data from the charging voltage curve as the model input to reduce the computational cost. Among all these input features used for model training, the direct external features are the easiest to obtain.

Due to the limited computational capability of the present BMSs, many features of the batteries are hard to obtain during the actual operation. A state monitoring method which can directly utilize the measurable features from the BMS for battery SOH estimation is highly desired. Ideally, the battery modeling and data pre-processing steps should be avoided to reduce computational efforts. Motivated by the correlation between the battery capacity and selected features of IC curves established in our previous work [8], we seek a method capable of estimating the battery capacity accurately by directly using partial charging curves without any pre-derivation or pre-smoothing steps. Driven by this purpose, this paper proposes a novel statistical learning method, random forest (RF) regression, to diagnose the SOH of LIB based on the voltage, current and time measurements during the charging process. The RF regression, initially presented by Breiman [20], is one of the most popular supervised machine learning algorithms and has been successfully applied to both classification and regression in many different fields, such as wind power forecast [21], wheat biomass estimation [22], and spatial prediction of soil organic carbon [23]. This method has been demonstrated to have the ability of well approximating variables with nonlinear relationships and also have high robustness performance against outliers. Despite the excellent predictive performance and reliable identification of relevant variables and interactions, few has employed the RF regression for SOH monitoring of lithium-ion batteries. The present work aims to fill this gap by proposing an RF regression-based estimation algorithm for online battery capacity estimation. This proposed approach has several salient characteristics desired for SOH estimation in BMSs. First, it is able to maintain high accuracy in the absence of any pre-selection of features, although confronted by significant noise in the predictive variables. Furthermore, while overfitting can cause inaccurate estimation with new testing data and thus negatively affect the model generality, the proposed algorithm is sufficiently robust against the overfitting phenomena.[1] In addition, compared to other machine learning techniques, e.g., ANN and SVM, it only needs a few tunable parameters and therefore requires low effort for offline model tuning [20].

The remainder of this article is organized as follows. Section 2 specifies the experiments, including implementation procedures, testing cells, and equipment. The proposed RF regression technique, feature selection, and model validation tools are presented in Section 3, followed by experimental implementation and discussion of the results in Section 4. A comparative study of the proposed SOH estimation algorithm and its two benchmarking methods is conducted in Section 5. Section 6 completes the present work with a concluding summary.

[1] A model that over-fits the data means that it is too flexible so that the isolated structures (e.g., noise) that are specific to the learning set can be captured erroneously.

2. Experiments and analysis

This section specifies the characteristics of experimental objects and testing equipment. The procedures of the cycling ageing test and the reference performance test are described in details.

2.1. Cell specifications and test equipment

The performance of the proposed method is evaluated based on data from commercial nickel-manganese-cobalt (NMC)/graphite pouch cells under different cycling conditions. Two types of NMC batteries from different manufacturers are considered: (i) seventeen type A cells are used for this research, each of which has a nominal capacity of 20 Ah; (ii) six type B cells, each with a nominal capacity of 31.5 Ah, are tested. The basic parameters of these battery cells are summarized in Table 1, where the manufacturers are not provided due to the project confidentiality.

Table 1
Specifications of tested NMC cells.

Properties	Type A cells	Type B cells
Manufacturers	A	B
Cell Weight	428 g	642 g
Nominal capacity (at 0.3I_t rate)	20 Ah	31.5 Ah
Nominal voltage (at 0.3I_t rate)	3.65 V	3.7 V
Energy density	174 Wh·kg^{-1}	180 Wh·kg^{-1}
Power density (at 50% SOC, 10 s)	2300 W·kg^{-1}	2300 W·kg^{-1}
Voltage range	3–4.15 V	3–4.2 V

During the experiments, all type A cells were placed in a climate chamber (350L CTS). The experiments were performed on a 16-channel, SBT0550L type battery test system manufactured by PEC with a current range of 0–50 A and a voltage range of 0–5 V. The voltage and current accuracy is in the full scale of ± 0.03%. The data acquisition system has a sampling frequency up to 100 Hz. All the type B cells were placed in an air-conditioned room within a temperature range of 25 ± 3 °C.

2.2. Testing procedures

Due to the specific application requirements, type A cells were cycled under more intense conditions than type B cells in terms of the current rate and the temperature. This section focuses on the testing procedures for type A cells. The cycling ageing matrix for type B cells has been presented in our previous work [8].

The batteries were aged under a charge/discharge test regime with constant currents. During charging, the current rate (C-rate) of $I_t/3$ was applied in the CC step until the battery reached its predefined higher voltage limit.[2] Then, the batteries were discharged immediately with the current rate of 1 I_t to the predefined lower voltage limit. Seventeen cells in total and divided into six groups were used for the cycling ageing test under different cycling depths of discharge (DOD), cycling middle-SOC (mid-SOC) levels and temperatures. The cells in groups 1, 2, 3 and 4 were cycled at a temperature of 35 °C under different cycling DOD (100%, 80% and 50%) and mid-SOC (50% and 65%). The cells in groups 5 and 6 were aged under an elevated temperature of 45 °C with different cycling DOD of 100% and 80%, respectively. The cycling conditions under which type A cells are tested further summarized in Table 2.

Table 2
Cycle ageing test matrix of type A cells.

	Cell No.	Cycling DOD, %	Mid-SOC, %	Temp, °C	Charge C-rate	Discharge C-rate
Group 1	1, 2, 3	100	50	35	$I_t/3$	1 I_t
Group 2	4, 5, 6	80	50	35	$I_t/3$	1 I_t
Group 3	7, 8, 9	50	50	35	$I_t/3$	1 I_t
Group 4	10, 11, 12	50	65	35	$I_t/3$	1 I_t
Group 5	13, 14, 15	100	50	45	$I_t/3$	1 I_t
Group 6	16, 17	80	50	45	$I_t/3$	1 I_t

Battery capacity tests were carried out through the whole experimental framework to acquire the information of battery health states. At each time to start capacity tests, the temperature of the climate chamber was adjusted to 25 °C. To do so, at least three hours of resting time were required for the batteries to cool down. Meanwhile, the constant current-constant voltage (CC-CV) mode was used during charging: the C-rate of $I_t/3$ was applied in the CC step until the battery reached its cut-off voltage of 4.15 V; this is followed by a CV step, in which the battery remained in the floating mode at the cut-off voltage until the current reached the minimum threshold of 0.1 I_t. After resting for 1 h, the battery was then discharged at a constant current of $I_t/3$ till 3.0 V. The obtained discharge capacity was used to calculate the battery SOH. Before the cycling tests to age battery cells, an initial capacity test was performed to determine each cell's actual rated capacity. Then, periodic capacity tests were performed with cycling intervals of every 100 full equivalent cycles (FECs) to acquire the trajectory of battery ageing. A summary of the complete test procedure is illustrated in Fig. 1.

3. Statistical learning approach

This section elaborates the feature extraction from real-time measured data for SOH estimation. Two characteristic features, indicative of the battery capacity fade, are exacted from the charging voltage-capacity curve and then used as the input features for model training and validating. Additionally, the fundamentals behind RF regression techniques as well as their performance evaluation tools are presented.

3.1. Feature extraction

The cells have been cycled to different SOH levels under varying cycling conditions. At various stages throughout the battery life, a capacity test with a full CC-CV charge cycle was applied on the cell with a fixed current at 25 °C, and the voltage vs. capacity ($V-Q$) data from this cycle was recorded. The actual battery capacity was obtained from the full charge/discharge cycle at this current rate, and the measured discharged capacity was used for calculating the real battery SOH. In reality, the battery discharge process is unpredictable as it depends on the varying utilization mode of the battery system. Nevertheless, the charge current is often constant in many applications such as charging of electric vehicles. Therefore, the charging $V-Q$ curves were chosen for battery SOH estimation.

A detailed flow diagram of feature extraction is presented in Fig. 2. A full charging $V-Q$ curve of a fresh NMC cell under constant current mode from the capacity test is used as a demonstration. The offline feature extraction process includes two steps:

Step 1: Defining voltage boundaries. In reality, batteries are rarely discharged to 0% SOC and charged to 100% SOC. Furthermore, the initial charging SOC usually is uncertain. Thus, the proposed SOH estimation method should be able to utilize partial charging curves. In this regard, only a specific voltage range of the testing data is selected

[2] I_t represents the current rate as documented in the standard IEC 61434 and defined by I_t = C/1 h [24].

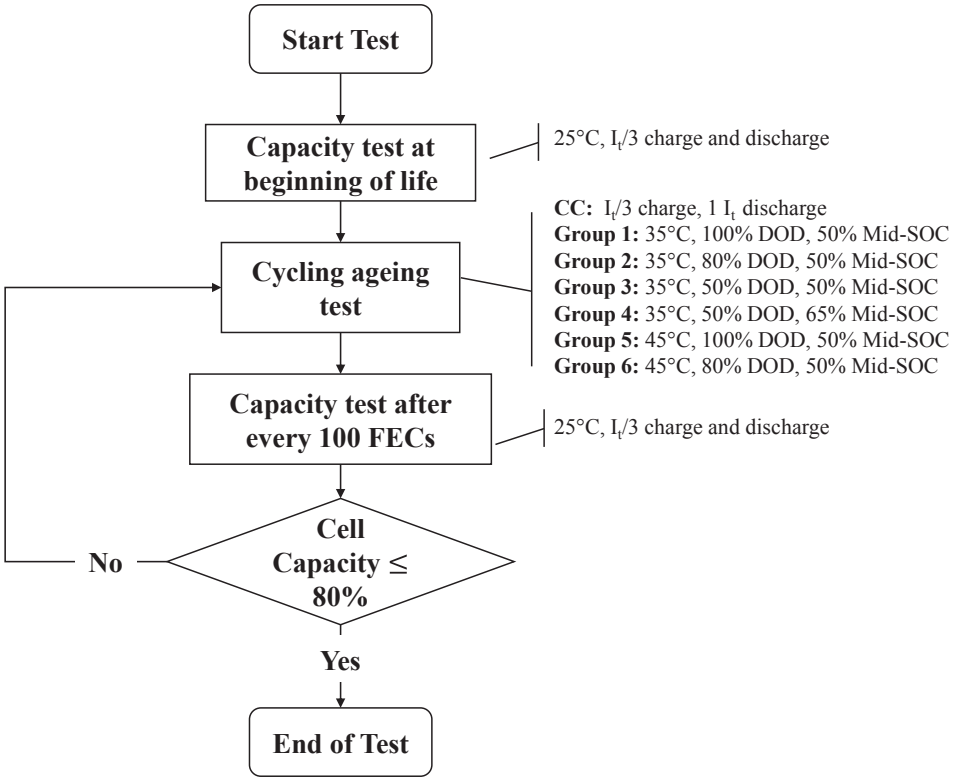

Fig. 1. Flow chart of the battery testing procedure for type A cells.

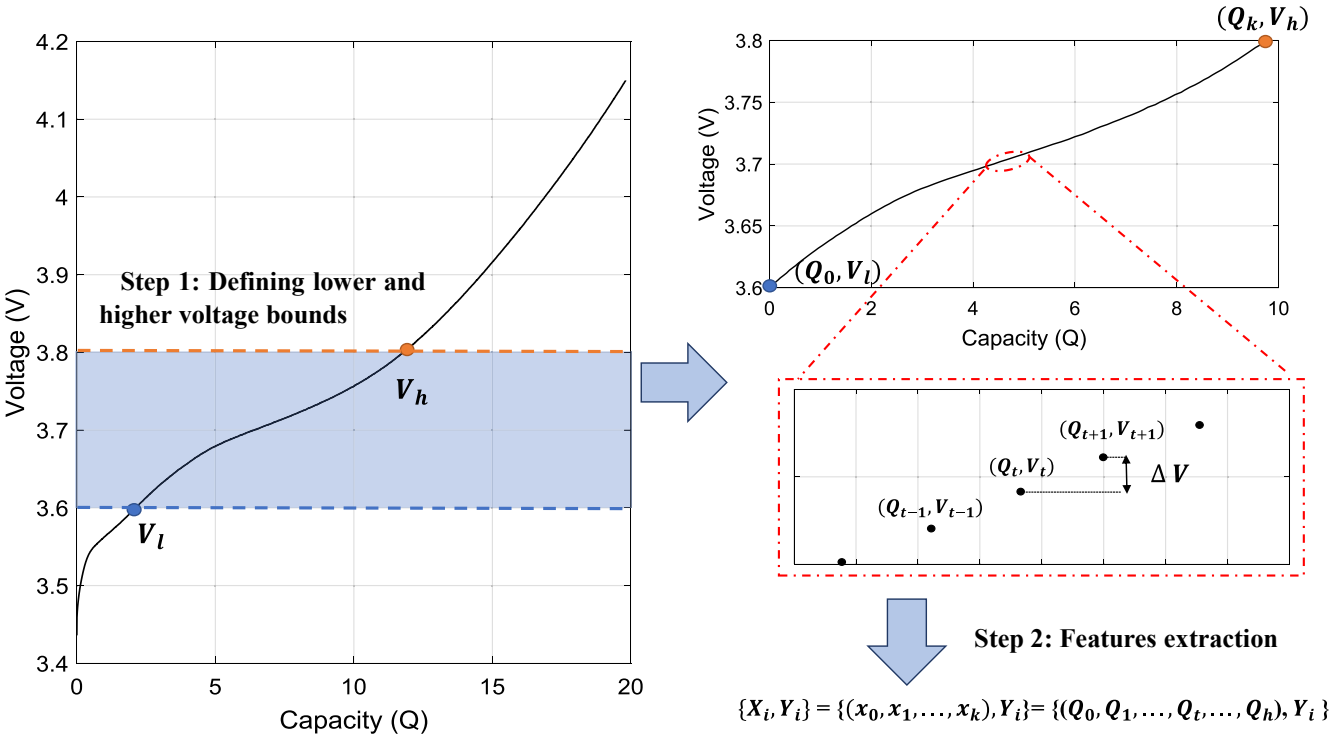

Fig. 2. Illustration of charge-related features extraction from a charge cycle.

for capacity estimation, and the voltage thresholds include the lower voltage bound V_l and upper voltage bound V_u.

Step 2: Feature extraction as input data. For each of the charging curves in the database obtained offline, the recorded relative capacity values (Q_t) with ΔV intervals in the specific voltage region (between V_l and V_u) is used as the input vector - X_i (also called predictor variables). Q_t is calculated based on coulomb counting by integrating the constant current with the time that battery charged from V_l to V_t, where V_t is the voltage at time t. The initial capacity Q_0 is therefore equal to 0. X_i is a vector with k variables, where $X_i = \{x_0, x_1, \ldots, x_t, \ldots, x_k\}$, and x_t is the relative charging capacity Q_t at time t. The value of k can be calculated by $k = (V_u - V_l)/\Delta V + 1$, which indicates the size of X_i. As the cell

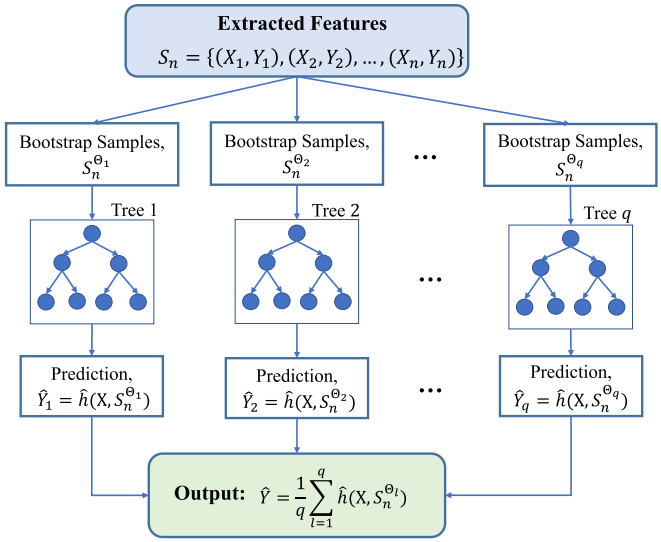

Fig. 3. Illustration of Random forest regression construction.

capacities for each charging curve in the offline database are known, each partial charging capacity vector X_i has an associated SOH value denoted as Y_i. The capacity test was carried out on cells after every 100 FECs, and the result consists of M constant charging curves throughout the ageing experiment. This indicates the cells have been cycled for $M \times 100$ FECs. Therefore, M feature vectors, $\{X_1, X_2, ..., X_M\}$, and their target SOH values, $\{Y_1, Y_2, ..., Y_M\}$, can be obtained from each of the training cells and form a training dataset $\{(X_i, Y_i),\ i = 1, 2, ..., M\}$. The training datasets are used to train and construct an RF model. For testing an RF regression model, the extracted feature vectors of the tested cells are fed into the trained model, which can produce estimated SOH values of different cells. For online application, when the battery voltage reaches V_l, the BMS will start to record the initial voltage V_l and capacity Q_0. When the battery voltage increases by ΔV, the corresponding charging capacity will be recorded until the upper voltage level V_u is reached. For example, when $V_l = 3.6$ V, $V_u = 3.8$ V and $\Delta V = 2$ mV are chosen, X would consist of the relative capacity values at the voltages of {3.600 V, 3.602 V, 3.604 V, ..., 3.800 V}. In short, the measured battery charge voltage and capacity are recorded in an onboard storage component of BMS as soon as the battery voltage reaches V_l, and the data recording process will be stopped when V_u is reached. The recorded datasets are then fed into the trained model for SOH estimation after the charging process.

3.2. Random forest regression

Random forest generates hundreds or even thousands of decision trees, which act as regression functions on their own, and the final output of the RF regression is the average of the outputs of all decision trees. A decision tree, also called Classification and Regression Tree (CART), is a statistical model firstly introduced by Breiman et al. in 1984 [25]. A decision tree is a nonparametric model. It does not assume any prior parameters for the class densities nor fixes the tree structure. The tree grows during the learning process depending on the complexity of the fed-in data [26]. Each decision tree consists of decision nodes and leaf nodes. The decision nodes evaluate each fed-in sample by a test function and pass it to different branches based on the features of the sample. Let X represent the input vector containing m features with $X = \{x_1, x_2, ..., x_m\}$, Y the output scalar and S_n the training set containing n observations which can be expressed as

$$S_n = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\},\ X \in \mathbb{R}^m,\ Y \in \mathbb{R}. \quad (1)$$

During training, the input data are split at each node by the algorithm, so that the parameters of split functions become optimized to fit with the S_n set. The decision tree has to make the best split among all variables during the first step. This splitting process starts at the root and each node applies its own split function to the new input X. This is repeated recursively until a terminal node (also called tree leaves) is hit. It is common to stop the tree when a maximum number of levels is reached, or when a node contains less than a predefined number of observations. At the end of this training process, a prediction function $\hat{h}(X, S_n)$ is constructed over S_n.

The random forest regression model is an extension of the CART technique and can offer better prediction performance. The training stage of RF is to construct multiple de-correlated decision trees. Each tree in RF is grown with a randomized subset of predictors and hence is called 'random' forest. RF uses L tree-structured base classifier $h(X,\ \Theta_k)$, where $k = 1,\ ...,\ L$, Θ_k is a family of independent, identically distributed random vectors, and X is an input vector [20]. It is an ensemble method that combines all the generated decision trees using an algorithm called 'bagging (or bootstrap aggregation)'. Bagging is a technique proposed by Breiman [27] and can be used with many regression methods to reduce the variance associated with prediction, thereby improving the prediction performance. RF can be built by randomly sampling a feature subset for each decision tree, and/or by randomly sampling a training data subset for each decision tree. This randomly collected sample process is called 'bootstrap', and a bootstrap sample is obtained by randomly selecting n observations with replacement from S_n, where each observation has the probability of $1/n$ to be selected. The bagging algorithm selects several bootstrap samples $(S_n^{\Theta_1},\ ...,\ S_n^{\Theta_q})$ and applies the previous tree decision algorithm to these samples in order to construct a collection of q prediction trees $\hat{h}(X, S_n^{\Theta_1}),\ ..., \hat{h}(X, S_n^{\Theta_q})$. The ensemble produces q outputs corresponding to each tree, $\hat{Y}_1 = \hat{h}(X,\ S_n^{\Theta_1}),\ \hat{Y}_2 = \hat{h}(X,\ S_n^{\Theta_2}),\ ...,\ \hat{Y}_q = \hat{h}(X,\ S_n^{\Theta_q})$.. Then the aggregation is performed by averaging the outputs of all trees. Consequently, the estimation $\hat{Y}$ of the output can be obtained by [28]

$$\hat{Y} = \frac{1}{q}\sum_{l=1}^{q} \hat{Y}_l = \frac{1}{q}\sum_{l=1}^{q} \hat{h}\left(X, S_n^{\Theta_l}\right), \quad (2)$$

where $\hat{Y}_l$ the output of l−th tree, and $l = 1,\ 2,\ ..., q$. The framework of using RF regression for prediction is illustrated in Fig. 3.

A prominent advantage of using bagging is to avoid the correlation of different trees, and the diversity of the trees can be accordingly increased by making them grow from different training data subsets created in RF [28]. Some data may be used more than once in the training, while others might never be used. Thus, greater stability of RF is achieved due to the utilization of bagging, which makes RF regression more robust when facing slight variations in input data [20]. Another advantage of bagging is the immunity to noise since it generates non-correlated trees through different training samples. A weak predictor may be sensitive to noise, but the average of several de-correlated decision trees can largely decrease the noise sensitivity [21]. An essential feature of RF is that the trees inside grow with no pruning, making them light from a computational perspective [28].

RF regression is very user-friendly with only two parameters to tune: the number of trees (n_{tree}) and the number of random features for each split (m_{try}) in the forest to build [29]. This is little need to fine-tune parameters to achieve excellent performance. In general, the more trees are grown in the forest, and the more robust and higher the accuracy of the prediction can be achieved. However, an increased amount of trees can lead to an enhanced computational burden. The generalization error converges as the number of trees increases, meaning that the estimation accuracy cannot be increased after reaching a certain point. The number of trees needs to be set sufficiently high to allow for this convergence, where the default value, $n_{tree} = 500$, is often used for prediction. m_{try} is a sensitive parameter determining model strength and defining the strength of each tree and the correlation between any two

Table 3
Construction of Random forest [29].

Step 1. Draw n_{tree} bootstrap samples from the original data, a bootstrap subset contains approximately 2/3 of the elements of the original dataset.
Step 2. To grow an unpruned regression tree using the bootstrap sample with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables.
Step 3. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, and average for regression).

trees in the forest. Increasing m_{try} can enhance the strength of each tree, however, the correlation between trees will be increased at the same time [30]. The improved tree strength improves model performance, while the increased correlation among trees can weaken it. It has been reported that the default value of m_{try}, one-third of the number of all predictor variables, is often a good choice [29]. To summarize, the basic steps for growing RF are listed in Table 3.

For each regression tree construction, a new training set (bootstrap samples) is generated with replacement from the original data. So, not all the samples are selected for the training of the q-th tree in the bagging process and some of the training data may be repeatedly used in the training sample. The samples that are not selected are included as part of another subset called out-of-bag (OOB) samples. Normally, two thirds of the new training samples are utilized for constructing the regression function whereas one third constitute the OOB sample. At each time when a regression tree is constructed with the training sample, the OOB sample is used to evaluate the performance of the regression tree. It is a kind of built-in cross validation process. In this way, RF can compute an unbiased estimation of the generalization error without using an external data subset as the trees have not seen these observations while training. Furthermore, the danger of overfitting can be largely reduced by using RF. This built-in validation feature improves the generalization capability of the RF. To obtain the total learning error, an average of the prediction error by each individual tree using the OOB sample can be obtained by

$$MSE \approx MSE^{OOB} = n^{-1}\sum_{i=1}^{n}[\hat{Y}(X_i) - Y_i]^2 \tag{3}$$

where $\hat{Y}(X_i)$ is the predicted output corresponding to a given input sample X_i, whereas Y_i is the observed output representing the real values and n is the total number of OOB samples. This error can determine how efficient the RF prediction is when it is exposed to unknown samples. The OOB error is an unbiased estimate of the generalization error, and it is proved that RF produces a limiting value of the generation error [20].

Matlab 2017a is used in this work to perform RF model training and testing. For the RF computations, we used the random forest package from Liaw et al. [29]. The package is in fact written in R, and a MEX interface is used in Matlab to call the C code used in the R package.

3.3. Performance evaluation tools

By comparing the actual SOH values from the experiment with the predicted ones, the prediction accuracy of the RF algorithm can be evaluated. Below, we list the metrics applied in this work for evaluating the quality of RF-generated predictions.

(1) Mean Absolute Error (MAE)

The MAE is defined by Eq. (4) and averages the absolute differences between the tested and predicted values. All the errors have the same weight in MAE. The smaller the MAE values, the more accurate the prediction result.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

where n represents the number of observations, y_i represents the experimental values and $\hat{y}_i$ represents the predicted values.

(2) Root Mean Square Error (RMSE)

The Root Mean-Square Error (RMSE) is frequently used to measure the difference between values predicted by the model and the observed values. It is very similar to MAE, but it penalizes larger absolute values by giving more weight to them than the MAE. The larger the difference between MAE and RMSE, the bigger the variance in the individual errors. RMSE is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{5}$$

(3) Maximum absolute error (ME)

The maximum absolute error the largest difference found between the estimated and observed values

$$ME = \max_{1\leqslant i\leqslant n}|y_i - \hat{y}_i| \tag{6}$$

(4) Goodness-of-fit (R^2)

R^2 is another measure of how closely the predicted values from a model match the observed values. The ideal R^2 value for a model is 1, which indicates that the model can explain all of the variability of the target class. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

where $\bar{y}$ represents the mean of response variables. When the predicted values are close to the experimental ones, the MAE, ME and RMSE are close to zero. Conversely, R^2 close to 1 indicates a good match between measured and predicted data.

3.4. Leave-one-out cross validation

Leave-one-out cross validation (LOOCV) is used here to assess the performance of the proposed RF regression for battery capacity estimation. Fig. 4 illustrates a schematic of the LOOCV approach used in this work. The complete feature dataset X consists of seventeen subsets, namely X_{c1}, X_{c2}, ..., X_{c17}, that were obtained respectively from the

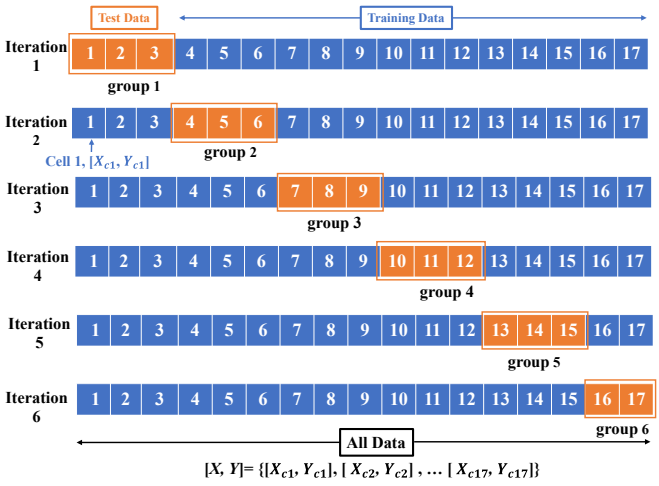

Fig. 4. A schematic display of leave-one-out cross validation process.

seventeen tested type A cells. Each subset is composed of M extracted features vectors, $\{X_1, X_2, ..., X_M\}$, obtained from M constant charging curves (see Section 3.1). As the seventeen cells are cycled under six different conditions, they are correspondingly divided into six independent data groups, as indicated in Table 2. The cells under the same cycling conditions retain similarities in the ageing pattern. We leave out the cells from one group as a test dataset (shown in orange) and compile the other five groups together to form a training set (shown in blue). Since the validation datasets are not used in the training process, the trained model can provide an approximately unbiased estimate. This process is repeated six times until each group is left out in turn. Here, 6-fold cross validation is performed for the proposed RF regression model. The LOOCV RMSE is computed as the average result of all six estimates, and the average represents the final error [31]. In this work, we estimate with the following equation:

$$LOOCV = \sqrt{\frac{\sum_{i=1}^{k} (RMSE_k)^2}{k}} \tag{8}$$

where k is the iteration counter during cross validation, and $RMSE_k$ is the $RMSE$ obtained from the k-th iteration according to Eq. (5).

4. Results on type A batteries

The data from the seventeen type A batteries aged under different cycling modes were used for verifying the effectiveness of the proposed method in capacity estimation. The influences of model parameter and input features on the estimation accuracy are discussed in this section.

4.1. Cycling ageing results

The battery capacity fades with cycling, and the correlation between the cycle number with the SOH evolution of type A cells is illustrated in Fig. 5. The cells cycled under the same conditions are marked with identical color. As can be seen, the capacity fade rates of cells under different cycling conditions are quite different. Specific stress factors can accelerate the ageing process, such as cycling DOD, temperature and mid-SOC. Cells in groups 1 (black), 2 (red) and 3 (green) were cycled under the same temperature and mid-SOC, but under different cycling DOD. Group 1 cells show the fastest ageing rate due to the deepest cycling DOD. At the elevated temperature of 45 °C, cells from group 5 aged faster than group 6 due to a higher cycling DOD. This observation is in line with the reports from the literature [32,33]. Temperature is another critical stress factor for battery capacity degradation. When the cycling temperature increased from 35 °C to 45 °C, the battery ageing rate increased significantly. One can observe this by comparing the capacity degradation data between cells from group 6 (magenta) and group 2 (red). The two groups of cells were cycled under the same DOD and middle SOC. Cells from group 6 suffered from a faster degradation rate due to the higher cycling temperature. These cells reached their EOL after only 900 FECs, while the SOH of cells from group 2 was around 95% after the same number of cycles. The impact of cycling middle SOC on battery degradation rate can also be observed by comparing the capacity fade between cells from group 3 (green) and group 4 (blue). During the first 1000 cycles, the two groups of cells show similar degradation behavior. After 1000 cycles, group 4 cells with a higher cycling middle SOC tend to degrade faster than group 3. Cells were not cycled to EOL due to time constraints. By comparing the capacity fade rates from the three groups, it is clear that increasing the battery cycling DOD, middle SOC and temperature can accelerate the battery capacity fade process. The full charge/discharge of batteries and high operating temperature should be avoided in real applications. The cycling data from seventeen type A cells will be used to verify the effectiveness of the proposed machine learning algorithm for capacity estimation in the next sub-sections.

4.2. Random forest regression model

In this section, the estimation performance of constructed RF regression model is examined. As mentioned in Section 3.1, characteristic features need to be selected first as input for the RF model. Herein, relative capacity values collected in the charging voltage region of 3.6–3.8 V in 2 mV intervals were used for model training and testing. To assess the quality of the model, the fit to the training data and the OOB validation statistics are considered, since RF does not overfit and a limiting value of the generalization error is obtained when the number of trees are increased. Fig. 6 presents the OOB error (MSE^{OOB}) as a function of the number of trees, where m_{try} was set as the default value (one-third of the total number of predictor variables). From approximately 200 trees onwards, the OOB error converge to a stable value for

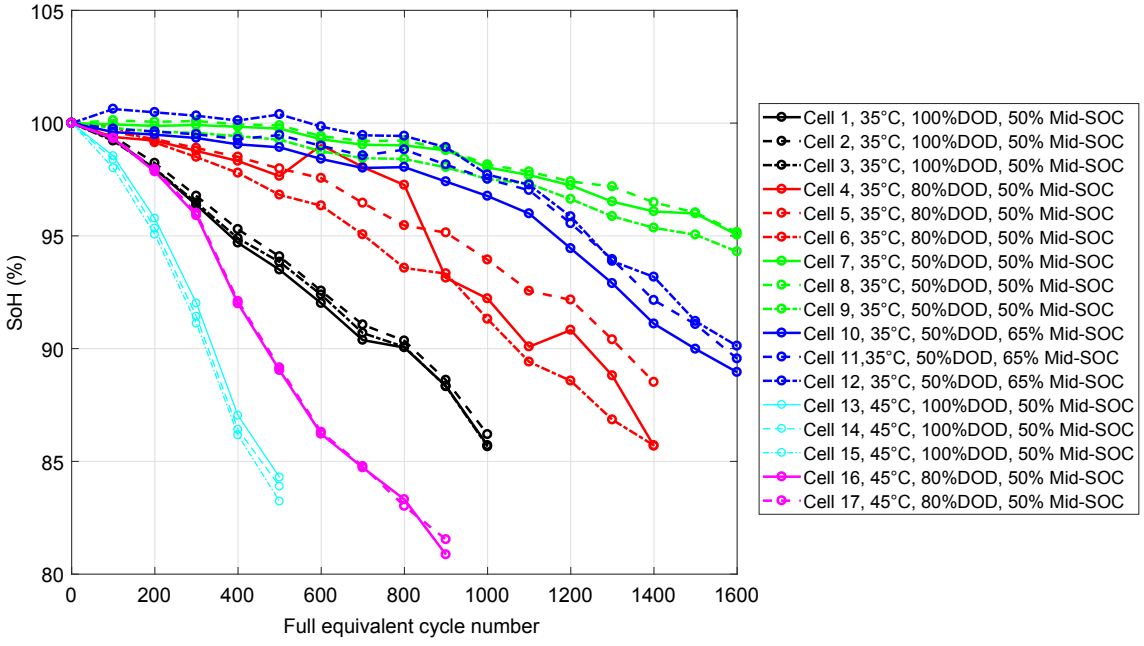

Fig. 5. SOH evolution of type A cells under different cycling conditions.

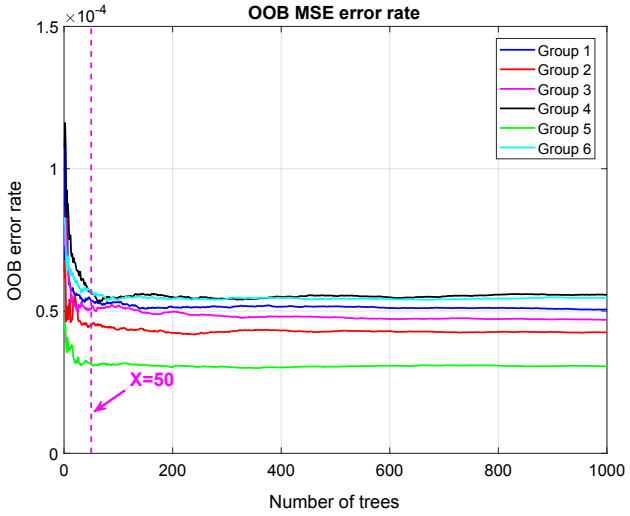

Fig. 6. OOB error converges rapidly as more trees are added to the random forest.

all tested groups. It is also found that increasing the number of trees above 200 does not improve the estimation accuracy but sacrifices the computational cost. From this figure, we can observe that the OOB error is a suitable estimator for error convergence detection and the random forest does not overfit as the value for OOB error converges to a small, but non-zero value. In this work, it was decided to construct the regression model with the default value of $n_{tree} = 500$. When the CPU time is limited for online application with massive datasets, a smaller value may be used for n_{tree} at the cost of some estimation accuracy for battery capacity.

The battery SOH estimation results from these datasets are listed in Table 4 (setting 2) and are further discussed in the next section. As can be seen, the trained RF model provides good estimation results as the RMSEs of all tested samples are less than 1.3%. These results demonstrate the ability of the RF model to learn the necessary information using the training set.

4.3. Influence of voltage range on estimation accuracy

In this work, the partial charge curve was generated from a full charge curve for battery capacity estimation. As described in Section 3.1, a specific charging voltage range should first be defined as the input vector X for training and testing. The capacity data were generated from a uniform distribution between a lower voltage bound V_l and an upper bound V_u with specific voltage intervals. Herein, we defined the voltage intervals as 2 mV for all tested cells. Three settings were considered for the voltage bounds: (i) Setting 1 - low initial SOC ($V_l = 3.4$ V and $V_u = 3.6$ V); (ii) Setting 2 - middle initial SOC ($V_l = 3.6$ V and $V_u = 3.8$ V); (iii) Setting 3 - high initial SOC ($V_l = 3.8$ V and $V_u = 4.0$ V).

The three settings correspond to three different SOC ranges: (i) 1–15% (iii) 15–60% and (iii) 60–85%, respectively. They were chosen to investigate how the voltage range affects the accuracy of battery capacity estimation. Fig. 7 illustrates the process of generating feature vectors from the charging capacity curves by different settings. The bottom three figures illustrate the partial voltage vs. relative capacity curves of cells after different cycles in the three different settings. The relative capacity is calculated based on Coulomb counting by integrating the charging current with time. The initial relative capacity values at the lower voltage limit are always zero for all three settings. For each of the settings, the feature data were generated using the charge curves obtained from the capacity test after every 100 FECs. The extracted partial charging curves were sampled and converted to feature vectors X, which correspond to the relative charge capacity calculated from the lower voltage bound with a specific voltage interval. The target displacement Y is the battery SOH calculated from the experiment. The M extracted feature vectors at different settings, $X_1, X_2, \ldots, X_M$, and their target capacity values, $Y_1, Y_2, \ldots, Y_M$, were used to train the RF regression model. Note that setting 1 represents the scenario where a cell undergoes the deepest discharge, compared to the scenario in settings 2 and 3, before the next charging cycle starts. Therefore, the data collected from setting 1 is the least common in real battery use, as batteries are barely discharged below 1% SOC. To obtain the data from settings 2 and 3, the starting charging voltage of the battery should be lower than 3.6 V and 3.8 V, respectively. The lower voltage limit requirements from settings 2 and 3 are suitable for most practical cases. The SOH estimation errors by the proposed RF regression method with the input data extracted from three different voltage ranges are compared in Table 4. For each group, the model was tested on the cells in this group and trained on all the other cells. The estimation performance of the proposed technique for each group was then measured by the four metrics: MAE, RMSE, maximal absolute error and R^2. Best values, namely lower errors and higher R^2, are marked with bold characters for more straightforward comparison. The overall performance of the developed model is evaluated by LOOCV RMSE, which is also presented in the same table. It is observed that the proposed method performs best with the input data from setting 2, with the lowest LOOCV RMSE of 0.82%. Therefore, to ensure the best SOH diagnosis accuracy, the data should be recorded in the middle SOC range. The possible reasons for this will be explained thoroughly in Section 5.1.1 by the IC analysis. Since the proposed RF regression model can provide the best estimation accuracy with the input data collected in the voltage region of 3.6–3.8 V, we define the middle SOC as the default voltage range for data collection in further discussion.

The SOH estimation results of the tested cells by the proposed methods with setting 2 are illustrated in Fig. 8(a)–(d). It is observed that the RF regression method can closely track the battery capacity fade trend throughout the cycling test for all tested cells with different ageing behavior. For instance, cells from group 5 were cycled under the harshest condition (45 °C and 100% DOD), and they showed a much faster capacity fade rate than the other cells. Nevertheless, the trained

Table 4
Errors of SOH estimations on the four groups of type A cells by the proposed RF method with the input from different voltage regions (Setting 1, 2 and 3).

Group	MAE (%)			RMSE(%)			Max Error(%)			R^2		
	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3
1	1.47	**0.67**	1.12	1.66	**0.74**	1.30	3.77	**1.45**	3.14	0.89	**0.97**	0.93
2	0.95	**0.62**	2.81	1.10	**0.82**	4.44	2.67	**2.47**	15.55	0.91	**0.97**	0.12
3	0.85	**0.36**	1.42	1.11	**0.48**	1.82	3.67	**2.22**	5.07	0.89	**0.98**	0.72
4	2.22	**0.62**	3.98	2.60	**0.84**	4.64	10.47	**2.67**	16.58	0.81	**0.98**	0.14
5	2.57	**1.23**	4.02	2.68	**1.26**	4.73	5.05	**2.67**	11.19	0.71	**0.95**	0.12
6	0.97	**0.43**	1.83	1.33	**0.55**	2.40	6.16	**1.50**	7.80	0.03	**0.92**	0.30
LOOCV	–	–	–	1.86	**0.82**	3.52	–	–	–	–	–	–

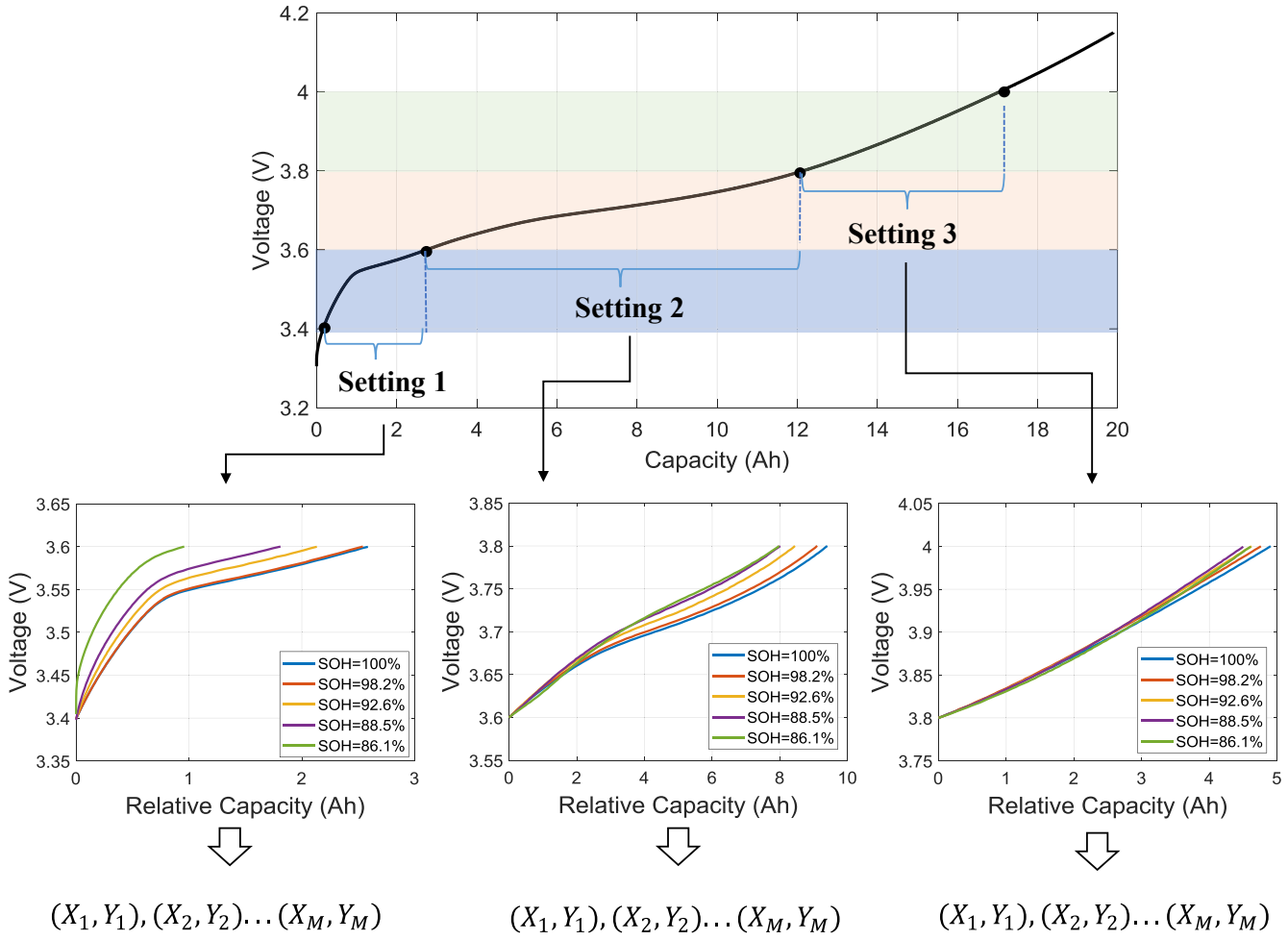

Fig. 7. Extraction of features from charging curves with three different voltage ranges. The extracted partial charging curves are sampled and converted to feature vectors X, which correspond to the relative charging capacity calculated from the lower voltage bond with specific voltage interval. The target displacement Y is the battery SOH calculated from the experiment.

RF regression model-based on the training dataset from the other groups is still able to give an accurate estimation result on these cells despite their different ageing patterns. This demonstrates the excellent estimation capability of the RF model for unknown datasets. However, it should be noted that the capacity tests on all tested cells were performed at 25 °C, and the influence of temperature on the input data is outside the scope of this work.

4.4. Influence of voltage intervals

In this section, a parametric study is carried out to investigate the effect of the sampling rate (voltage intervals) on the estimation accuracy of the RF regression model. For each tested battery, we collected several dataset with different voltage intervals (2, 5, 7, 10, 15, 30, 50 and 100 mV). Increased voltage interval corresponds to decreased sampling rate and the number of input features. For instance, when the cells are charged from 3.6 V to 3.8 V with 2 mV interval, totally 101 relative charge capacity values can be recorded by the BMS. If the voltage interval increases to 100 mV, only 3 data points can be recorded during the charging process. A plot of the RMSE for the estimation output from RF model with different sampling rates is shown in Fig. 9. It can be observed that for all tested groups the RMSEs are almost constant from 2 to 30 mV, while a sudden increase of RMSE can be found for the cases with voltage intervals of more than 30 mV. The results suggest that the sampling rate has a great impact on the prediction accuracy and it should be kept in an appropriate range. When the voltage interval is less than 30 mV, the proposed RF regression model produces relatively small errors for all tested cells.

Increasing the sampling rate requires enhanced resolution of data recording as well as the memory of a BMS. Hence, the most appropriate sampling rate should be as low as possible while not compromising the estimation accuracy of the model. A voltage interval of 30 mV is therefore recommended for the online application of the proposed method. With this setting, the minimum time interval between two adjacent data points is around 480s for a fresh cell under charging current rate of $I_t/3$. Given the sampling rate of most commercial BMS can reach 1 Hz, the above specification is good enough to capture characteristics of the signals. Moreover, only 7 data points are collected by using a voltage interval of 30 mV, and the decreased amount of input features can largely enhance the computational speed of the proposed method for online SOH estimation.

5. Discussion

The use of RF regression for improving the accuracy and computational efficiency in the battery capacity estimation over existing methods is studied. To this, we compared it with two commonly used data-driven techniques, the IC analysis and Gaussian Process regression. The two techniques were chosen due to their simplicity in processing the input data. They resemble the proposed regression method, as only partial charging $V-Q$ data are required for SOH monitoring. The proposed model was also tested and validated on a different type of NMC

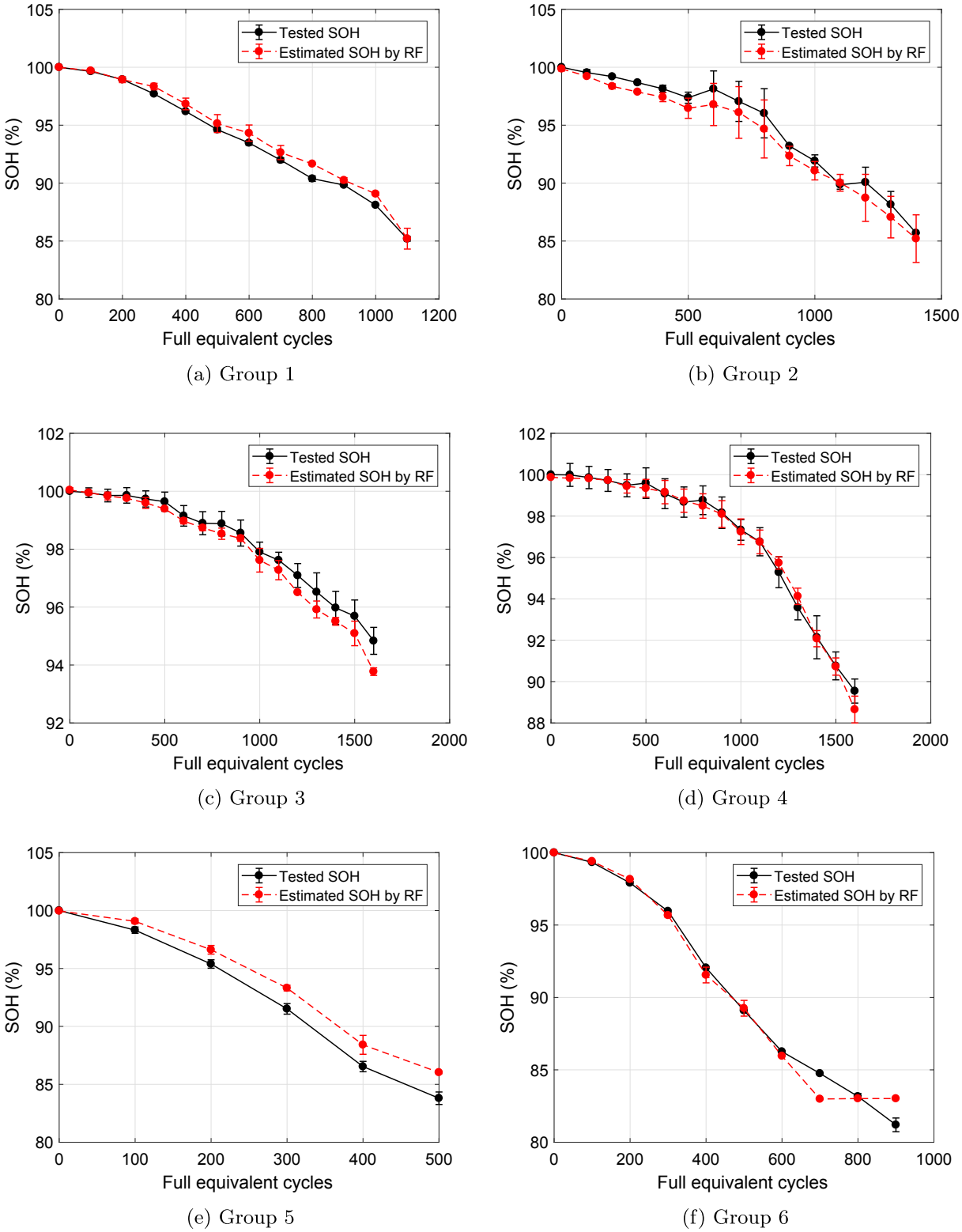

Fig. 8. SOH estimation results of tested cells from (a) group 1 (b) group 2 (c) group 3 (d) group 4 (e) group 5 and (f) group 6 by the proposed RF method. The estimated SOH data for the cells from the same group are represented as an average with periodic error bars showing ± one standard deviation.

batteries, and the performance the model is discussed in this section.

5.1. Comparative study with existing SOH estimation techniques

5.1.1. Comparison with IC analysis

As mentioned in the introduction part, IC analysis is an efficient and quick tool for on-line estimation of battery SOH. Incremental capacity is calculated by differentiating the change in battery capacity to the change in terminal voltage during either charging or discharging, as mathematically given by

$$\frac{dQ}{dV} \approx \frac{\Delta Q}{\Delta V} = \frac{Q_t - Q_{t-1}}{V_t - V_{t-1}}. \quad (9)$$

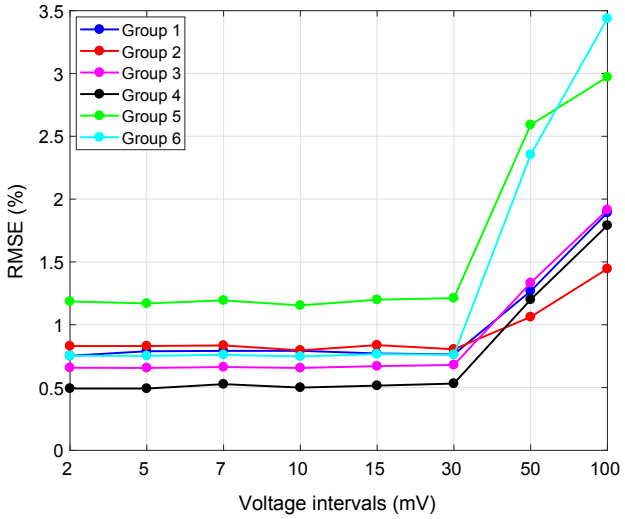

Fig. 9. RMSE for the input data with different voltage intervals from the type A cells.

The IC curves of the tested cells generated from constant current charging curves with a current rate of $I_t/3$ were used for battery SOH estimation. Here, a Gaussian filter was applied to filter out the noise on IC curves according to our previous work [8]. The obtained results are illustrated in Fig. 10(a). Two prominent peaks around 3.57 V and 3.70 V can be observed on the IC curves of fresh batteries and are named as peak A and peak B, respectively. According to the previous research, peak A is correlated with the reactions on the negative electrode [34,35], and peak B is likely a result of the sequential phase transition process in the NMC cathode [36,37]. The area under each IC peak represents the capacity of the related reaction during the charging process [38].

It is observed that the two peaks shift towards higher voltage levels. The height of peak B decreases with the battery capacity fade, and it shows a linear relationship with the battery SOH as illustrated in Fig. 10(b). The estimation function for tested batteries can be expressed as

$$SOH\,(\%) = 0.574 \times PH_B + 56.01 \tag{10}$$

where PH_B indicates the height of peak B. The estimation results of the developed linear function for battery SOH estimation are listed in Table 5. With the developed linear equation, the SOH of batteries can be estimated with less than a maximum error 4% and an RMSE of 1.59%. The estimation is the poorest in the case of group 5 cells which are cycled under elevated temperature (45 °C) and 100% DOD. It suggests that the estimation accuracy of IC analysis is influenced by the cycling conditions. However, a detailed study of temperature effects is beyond the scope of this paper. IC analysis is not a machine learning technique, and the LOOCV RMSE is therefore not used for evaluating its general performance. Instead, the estimation accuracy is compared for each group of cells between IC analysis and RF regression. Several important observations can be made based on the results:

- Both IC and RF provide reasonable SOH estimates, but RF outperforms IC with better estimation accuracy in terms of RMSE for all tested cells. This suggests the robustness of the proposed method for capacity estimation of cells under different cycling condition and with different ageing behavior.
- IC curves are sensitive to the noise in the $V-Q$ curves. A proper smoothing technique to provide proper noise filtering and preserve the important features on the curves is crucial. In contrast to the above, filtering is unnecessary for RF regression, since this technique is highly robust to noise in predictors and thus does not require a pre-selection of variables.
- It is observed from Fig. 10(a) that the height of peak B decreases significantly with battery ageing. As the area under the IC peak represents the capacity of the related reaction during the charging process, it is evident that the area variation of peak B contributes the most to the total battery capacity fade. Peak B is located in the voltage region of 3.6–3.8 V and this explains why the input datasets from setting 2 give the best estimation result in the RF regression model. No noticeable feature can be found in the IC curve in the voltage region of 3.8–4.0 V, which can explain why the proposed RF model produces the worst results in setting 3. Apparently, the performance of RF model is dependent on how strongly the partial $V-Q$ data is correlated with cell capacity fade.
- IC analysis is a powerful tool for understanding the battery degradation behavior and can be used to aid the feature selection for a black-box model such as RF. When applying the RF regression model on different battery chemistry, the IC technique can be utilized for finding the part of $V-Q$ data which contributes most to the whole cell capacity. In this way, trial and error in selecting the region of interests on the charging curve can be avoided during the feature selection process.

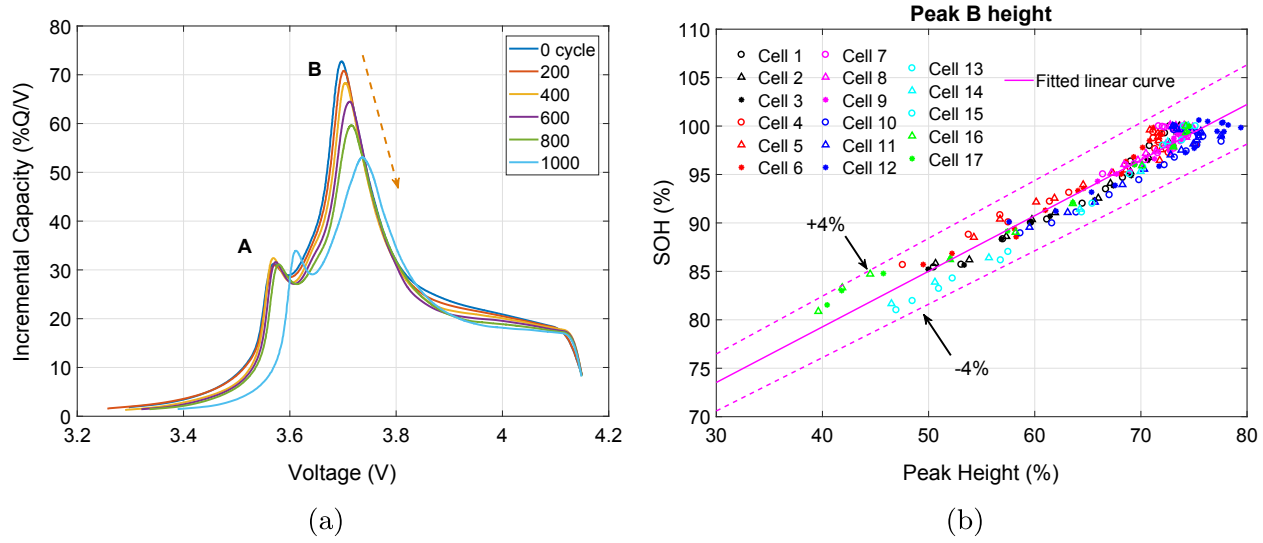

Fig. 10. (a) IC curves of cell 1 from type A cells at different ageing states (b) Correlation between measured battery SOH and normalized peak B height.

Table 5
Estimation error comparison among the proposed RF regression with IC analysis and GR regression method.

Group	MAE (%)			RMSE(%)			Max Error(%)			R^2		
	IC	GR	RF	IC	GR	RF	IC	GR	RF	IC	GR	RF
1	0.88	1.02	**0.67**	0.93	1.11	**0.74**	1.98	2.31	**1.45**	0.97	0.93	**0.97**
2	1.29	0.91	**0.62**	1.43	1.08	**0.82**	2.98	2.97	**2.47**	0.95	0.95	**0.97**
3	1.24	0.57	**0.36**	1.32	0.76	**0.48**	2.52	3.52	**2.22**	0.88	0.96	**0.98**
4	1.34	0.66	**0.62**	1.59	**0.84**	**0.84**	2.36	**2.17**	2.67	**0.98**	**0.98**	**0.98**
5	1.48	1.83	**1.23**	1.44	2.06	**1.26**	2.83	5.21	**2.67**	**0.98**	0.81	0.95
6	0.91	**0.36**	0.43	1.16	**0.47**	0.55	2.85	1.52	**1.50**	0.77	**0.92**	**0.92**
LOOCV	–	–	–	–	1.16	**0.82**	–	–	–	–	–	–

5.1.2. Comparison with Gaussian Process (GP) regression

GP regression is a machine learning technique and has been employed for estimating battery capacity with partial charging curves. The details of this approach can be found in the work of Richardson et al. [15], which is similar to our approach by utilizing the input data of capacity recorded at equispaced voltage points in a specific voltage range. However, they need to apply a pre-smoothing step for the input data with Savitzky-Golay filtering to improve the ratio of signal to noise. Additionally, the voltage intervals used in the article of [15] are much larger than in this work, which means fewer input datasets for model training and testing. In order to compare the performance of the proposed technique with GP regression in terms of estimation accuracy and computational cost, the input data used for the two techniques should be the same. Therefore, such a data smoothing step was not adopted here and the raw data of setting two (relative capacity recorded in the voltage range of 3.6–3.8 V with 2 mV intervals) were used for the comparison of the two methods. The Gaussian process regression was constructed with a Matern (5/2) kernel function to achieve the mapping from inputs X to outputs Y. The GP regression model was built in Matlab environment with the 'fitrgp' function. A more detailed process of setting up this model can be found in https://nl.mathworks.com/help/stats/fitrgp.html. The estimation errors are listed and compared with RF regression in Table 6. LOOCV was used to validate the performance of GP regression. The LOOCV RMSE of RF regression is 0.82%, while the value of GP regression is 1.16%. Both GP and RF regression provide good estimation results, although RF shows a slightly better estimation accuracy.

The computational cost is an important factor in evaluating the method from a practical perspective. The algorithms were repeatedly executed for ten times to obtain the average computational time. Both of the regression models were performed in MATLAB R2017a (64bit) and operated on a computer with 2.90 GHz processor and 16.0 GB RAM. The results of computational time for executing the two models are shown in Table 6. The computational time for all tested cell groups is compared. It is clear that RF regression is an order of magnitude faster than GP regression. This demonstrates the superior computational efficiency of the proposed RF regression method and shows great potential for real applications, particularly where the computational cost is a constraint.

Table 6
Computational time of the proposed method and Gaussian process regression.

Group	Computational time (s)	
	GP regression	RF regression
1	23.62	1.39
2	11.32	1.28
3	15.80	1.34
4	17.56	1.05
5	15.56	1.37
6	22.65	0.79

The proposed method is compared with two existing data-driven SOH estimation techniques, IC analysis and GP regression, in terms of estimation accuracy and computation cost. All of them can track the battery SOH evolution well with partial $V-Q$ data. RF regression outperforms the other two methods with the tested input data. To obtain IC curves, derivation and filtering steps are required to get smooth IC curves. Then the features on the IC curves need to be identified for SOH estimation with the proposed regression function. When using the machine-learning based methods of GP regression and RF regression, such pre-manipulating steps of the $V-Q$ data can be avoided. The raw charging capacity data can be fed directly into the proposed model for training and testing. Nevertheless, RF regression provides a better estimation effect than GP regression due to the smaller LOOCV RMSE and faster computation speed.

5.2. Testing and validation on type B batteries

The RF regression method was tested and validated on type B cells with a nominal capacity of 31.5 Ah from a different battery manufacturer. Six cells were all cycled at room temperature for the duration of the experiments. The cells were divided into 3 groups, with each group undergoing a different cycling DOD as described in our previous publication [8]. All the cells were cycled at room temperature with a charge/discharge current rate of $I_t/2$. In all cases, characterization test was periodically carried out after every 100 FECs, whereby a $I_t/3$ charge–discharge cycle was applied. The calculated SOH for the cells from all three groups are plotted against the FECs in Fig. 11(a). As type B cells have different features from type A cells in terms of capacity, maximum voltage limit, manufacturing process, etc., the developed RF regression model trained on type A cells cannot be directly applied on type B cells. Therefore, we used the same procedures described in the previous section to train the RF regression model on type B cells.

Feature selection for type B cells follow the same process as described in Section 3.1. The charge capacity values were recorded in the voltage range of 3.6–3.8 V with 2 mV intervals. The SOH estimation results on the three groups of type B cells are graphically summarized in Fig. 11(b)–(d), respectively. The estimated SOH values closely match the test results, and the estimation errors are listed in Table 7. A comparison of the results between type A and B cells suggests that the RF method works better on type B cells. The LOOCV RMSE of type A cells is 0.84% while the LOOCV RMSE of type B cells is less than 0.31%. The increased accuracy for type B cells can likely be attributed to the fact that they exhibit a somewhat more homogeneous capacity fade behavior throughout the cycling test, as can be seen from Fig. 11(a). Conversely, the differences in the SOH evolution of type A cells between the cells cycled under different condition are much larger. Even the cells cycled under the same conditions can exhibit quite different ageing behaviors, such as the cells in group 2. The SOH estimation accuracy on type B cells further emphasises that the proposed method is effective in estimating the SOH of NMC LIBs, regardless of their chemical features or the manufacturing process.

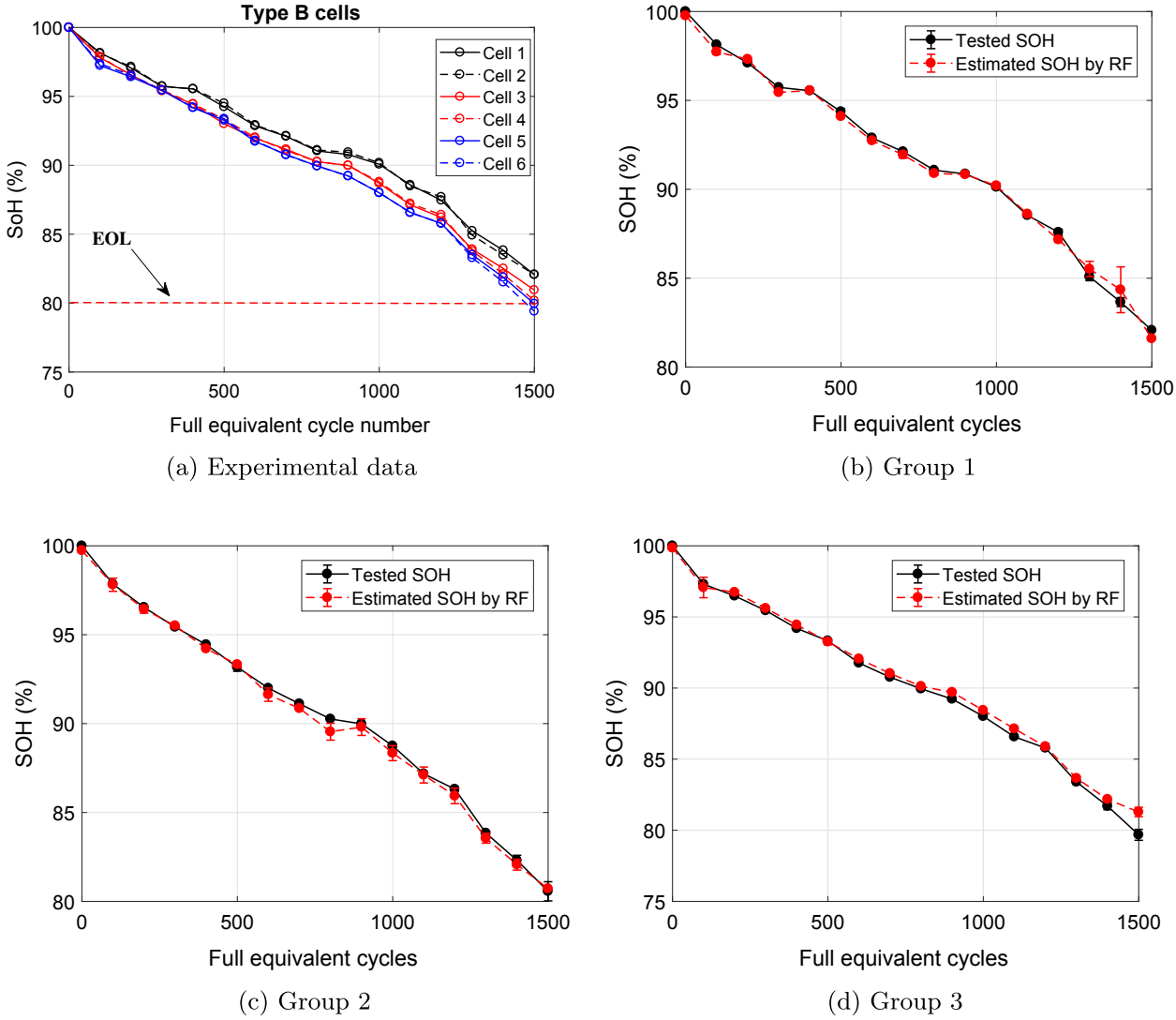

Fig. 11. (a) SOH evolution of all tested cells from type B cells; battery SOH estimation results of type B cells from (b) group 1 (c) group 2 and (d) group 3 with proposed RF regression method. The estimated SOH data for the cells from the same group are represented as an average with periodic error bars showing ± one standard deviation.

Table 7
SOH estimations on tested type B cells with proposed RF methods.

Group	MAE (%)	RMSE(%)	Max Error(%)	R^2
1	0.30	0.39	1.70	0.99
2	0.34	0.38	1.16	0.99
3	0.44	0.53	2.66	0.99
LOOCV	–	0.31	–	–

5.3. Challenges and future work

This paper presents an accurate online capacity estimation technique for NMC batteries. One hurdle to the practical implementation of this method is the variable operating temperature that have a significant impact on battery $V-Q$ curves. Although the present results are derived based on the capacity tests at a constant temperature of 25 °C, the method should be applicable to varying temperature conditions when enough training data are available. The most desirable solution for SOH estimation lies in developing a universal tool, which is applicable to any type of batteries with good accuracy under various conditions. Future work will be focused on validating the proposed method with different battery chemistries at varying operating temperatures and current rates. To achieve this goal, a large battery cycle test matrix at different constant stress conditions such as temperature, current rate and discharging DOD will be generated.

The machine learning technique proposed by this work is not aimed to offer electrochemical insight into the battery degradation process. Understanding the underlying ageing mechanisms can help in finding the most sensitive features for battery capacity fade at various operating conditions, and then potentially improves the SOH estimation performance. The ageing process often involves complex and coupled physical–chemical processes in complicated operating conditions. To understand the battery degradation mechanisms, post-mortem analysis using destructive methods like X-ray diffraction and scanning electron microscope can be helpful. Additionally, the adaptivity of the developed RF regression method for online application will be evaluated by a hardware-in-the-loop (HIL) test bench.

6. Conclusions

This work presents a novel machine learning approach for online battery capacity estimation. The random forest regression model is developed to approximate the relationship between characteristic

features extracted from the charging voltage-capacity curve and capacity. The charging capacity data recorded in a specific voltage region with certain intervals is used as input data for model training and testing. The incremental capacity analysis is used as a complementary method for selecting the input features, as the evolution of peaks in incremental capacity curves can directly indicate which part of the voltage-capacity data is strongly correlated with battery capacity fade. The proposed algorithm can accurately estimate the capacity of batteries aged under different cycling conditions with a maximum root-mean-square error of 1.3%. Furthermore, with this method, the online recorded data from the battery management system can be directly used as the input feature vectors for monitoring battery health without any pre-processing procedures, e.g. smoothing or derivation. As a result, this algorithm has a low computational cost but does not compromise the ability to process a large amount of data. These advantages make the proposed algorithm well suitable for online monitoring of battery state of health.

Acknowledgment

The authors would like to thank ENGIE Laborelec for supporting this research work and also thank Flanders Make for supporting the battery team in Vrije Universiteit Brussel. Y. Li would like to thank Andrej Turk from the University of Cambridge for helpful discussions and proofreading.

References

[1] Plett GL. Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part 3. State and parameter estimation. J Power Sources 2004;134:277–92.
[2] Zou C, Manzie C, Nešić D, Kallapur AG. Multi-time-scale observer design for state-of-charge and state-of-health of a lithium-ion battery. J Power Sources 2016;335:121–30.
[3] Schwunk S, Armbruster N, Straub S, Kehl J, Vetter M. Particle filter for state of charge and state of health estimation for lithium–iron phosphate batteries. J Power Sources 2013;239:705–10.
[4] Li S, Pischinger S, He C, Liang L, Stapelbroek M. A comparative study of model-based capacity estimation algorithms in dual estimation frameworks for lithium-ion batteries under an accelerated aging test. Appl Energy 2018;212:1522–36.
[5] Tang X, Zou C, Yao K, Chen G, Liu B, He Z, et al. A fast estimation algorithm for lithium-ion battery state of health. J Power Sources 2018;396:453–8.
[6] Dubarry M, Liaw BY. Identify capacity fading mechanism in a commercial $LiFePO_4$ cell. J Power Sources 2009;194:541–9.
[7] Weng C, Cui Y, Sun J, Peng H. On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression. J Power Sources 2013;235:36–44.
[8] Li Y, Abdel-Monem M, Gopalakrishnan R, Berecibar M, Nanini-Maury E, Omar N, et al. A quick on-line state of health estimation method for Li-ion battery with incremental capacity curves processed by gaussian filter. J Power Sources 2018;373:40–53.
[9] Wang L, Zhao X, Liu L, Pan C. State of health estimation of battery modules via differential voltage analysis with local data symmetry method. Electrochim Acta 2017;256:81–9.
[10] Wu J, Wang Y, Zhang X, Chen Z. A novel state of health estimation method of Li-ion battery using group method of data handling. J Power Sources 2016;327:457–64.
[11] You G-w, Park S, Oh D. Real-time state-of-health estimation for electric vehicle batteries: A data-driven approach. Appl Energy 2016;176:92–103.
[12] Klass V, Behm M, Lindbergh G. A support vector machine-based state-of-health estimation method for lithium-ion batteries under electric vehicle operation. J Power Sources 2014;270:262–72.
[13] Nuhic A, Terzimehic T, Soczka-Guth T, Buchholz M, Dietmayer K. Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods. J Power Sources 2013;239:680–8.
[14] Hu C, Jain G, Schmidt C, Strief C, Sullivan M. Online estimation of lithium-ion battery capacity using sparse bayesian learning. J Power Sources 2015;289:105–13.
[15] Richardson RR, Birkl CR, Osborne MA, Howey D. Gaussian process regression for in-situ capacity estimation of lithium-ion batteries. IEEE Trans Ind Info, to be published, https://doi.org/10.1109/TII.2018.2794997.
[16] Yang D, Zhang X, Pan R, Wang Y, Chen Z. A novel gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. J Power Sources 2018;384:387–95.
[17] Pan H, Lü Z, Wang H, Wei H, Chen L. Novel battery state-of-health online estimation method using multiple health indicators and an extreme learning machine. Energy 2018;160:466–77.
[18] Wang Z, Ma J, Zhang L. State-of-health estimation for lithium-ion batteries based on the multi-island genetic algorithm and the gaussian process regression. IEEE Access 2017;5:21286–95.
[19] Berecibar M, Devriendt F, Dubarry M, Villarreal I, Omar N, Verbeke W, et al. Online state of health estimation on nmc cells based on predictive analytics. J Power Sources 2016;320:239–50.
[20] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[21] Lahouar A, Slama JBH. Hour-ahead wind power forecast based on random forests. Renew Energy 2017;109:529–41.
[22] Wang L, Zhou X, Zhu X, Dong Z, Guo W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J 2016;4:212–9.
[23] da Silva Chagas C, de Carvalho Junior W, Bhering SB, Calderano Filho B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 2016;139:232–40.
[24] IEC61434:1996. Secondary cells and batteries containing alkaline or other non-acid electrolytes – guide to designation of current in alkaline secondary cell and battery standards. International Standard 1996-10-03.
[25] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC Press; 1984.
[26] Alpaydin E. Introduction to machine learning. MIT Press; 2014.
[27] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.
[28] Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev 2015;71:804–18.
[29] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2:18–22.
[30] Peters J, De Baets B, Verhoest NE, Samson R, Degroeve S, De Becker P, et al. Random forests as a tool for ecohydrological distribution modelling. Ecol Model 2007;207:304–18.
[31] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning vol. 112. Springer; 2013.
[32] de Hoog J, Timmermans J-M, Ioan-Stroe D, Swierczynski M, Jaguemont J, Goutam S, et al. Combined cycling and calendar capacity fade modeling of a Nickel-Manganese-Cobalt Oxide cell with real-life profile validation. Appl Energy 2017;200:47–61.
[33] Sarasketa-Zabala E, Gandiaga I, Martinez-Laserna E, Rodriguez-Martinez L, Villarreal I. Cycle ageing analysis of a $LiFePO_4$/graphite cell with dynamic model validations: Towards realistic lifetime predictions. J Power Sources 2015;275:573–87.
[34] Dubarry M, Devie A, Liaw BY. The value of battery diagnostics and prognostics. J Energy Power Sources 2014;1:242–9.
[35] Jung R, Metzger M, Maglia F, Stinner C, Gasteiger HA. Oxygen release and its effect on the cycling stability of $LiNi_xMn_yCo_zO_2$(NMC) cathode materials for Li-Ion batteries. J Electrochem Soc 2017;164:A1361–77.
[36] Ryu W-H, Lim S-J, Kim W-K, Kwon H. 3¯D dumbbell-like $LiNi_{1/3}Co_{1/3}Mn_{1/3}O_2$ cathode materials assembled with nano-building blocks for lithium-ion batteries. J Power Sources 2014;257:186–91.
[37] Xu J, Chou S-L, Gu Q-f, Liu H-K, Dou S-X. The effect of different binders on electrochemical properties of $LiNi_{1/3}Co_{1/3}Mn_{1/3}O_2$ cathode material in lithium ion batteries. J Power Sources 2013;225:172–8.
[38] Han X, Ouyang M, Lu L, Li J, Zheng Y, Li Z. A comparative study of commercial lithium ion battery cycle life in electrical vehicle: Aging mechanism identification. J Power Sources 2014;251:38–54.